Chapter 2
Environment-Related Robustness Issues

In practical applications, many environment-related factors may influence the performance of speaker recognition. There is often *no prior* knowledge of these factors in advance, which makes the environment-related robustness issue more difficulty. In this chapter, three environment-related factors, background noise, cross channel and multiple-speaker, are summarized and their corresponding robustness issues are discussed.

2.1 Background Noise

The speech wave recorded in real environments often contains different types of background noises such as white noise, car noise, music etc. The background noise has adverse impact on speaker modeling and disturbs the evaluation testing, and so degrades the performance of speaker recognition system. The research on background noise robustness generally has four directions: speech enhancement, feature compensation, robust modeling, and score normalization.

2.1.1 Speech Enhancement

Despite the fact that the conventional and the state-of-the-art speech enhancement techniques have been gained satisfactory effects, employing signal-level enhancement has shown to be effective in improving speaker recognition in noisy environments. In [1], the subtractive noise suppression analysis was presented, and the spectral subtraction was proposed to suppress stationary noise from speech by subtracting the spectral noise bias calculated during non-speech activity and attenuate the residual noise left after subtraction. Since this algorithm resynthesizes a speech waveform, it can be used as pre-processing to a speaker recognition

T.F. Zheng and L. Li, *Robustness-Related Issues in Speaker Recognition*,
SpringerBriefs in Signal Processing, DOI 10.1007/978-981-10-3238-7_2

system. In [2], different techniques were used to remove the effect of additive noise on the vocal source features WOCOR and the vocal track features MFCC. And a frequency-domain approach was proposed to denoise the residual signal and hence improve the noise-robustness of WOCOR. However, these methods do not perform well when noise is nonstationary. RASTA filtering [3] and cepstral mean normalization (CMN) [4] have been used in speaker recognition but they are mainly intended for convolutive noises. Inspired by auditory perception, computational auditory scene analysis (CASA) [5] typically segregates speech by producing a binary time-frequency mask. To deal with noisy speech, Zhao applied CASA separation and then reconstructed or marginalized corrupted components indicated by a CASA mask. It was further shown in [6] that these algorithms might either enhance or degrade the recognition performance depending on the noise type and the SNR level.

2.1.2 Feature Compensation

There are also algorithms to improve the system robustness in feature domain. In [7], 12 different short-term spectrum estimation methods were compared for speaker verification under the additive noise contamination. Experimental results conducted on the NIST 2002 SRE show that the spectrum estimation method has a large effect on recognition performance and the stabilized weighted LP (SWLP) and the minimum variance distortionless response (MVDR) methods can yield approximately 7 and 8% relative improvements over the standard DFT method in terms of EER. Lei et al. [8, 9] proposed a vector Taylor series (VTS) based i-vector model for noise-robust speaker recognition by extracting synthesized clean i-vectors to be used in the standard system back-end. This approach brought significant improvements in accuracy for noisy speech conditions. Martinez et al. [10] tried to model non-linear distortions in cepstral domain based on a nonlinear noise model in order to relate clean and noisy cepstral coefficients and help estimate a "cleaned-up" version of i-vectors. Moreover, to avoid the high computational load of the i-vector modelling in the proposed noisy environment, a simplified version is followed, where the sufficient statistics are normalized with their corresponding utterance-dependent noise adapted UBM.

2.1.3 Robust Modeling

The research on model robustness against noise usually adopts model compensation algorithms to decrease the mismatch between the test and the training utterances.

The parallel model combination (PMC) was first introduced in speech recognition [11] in advance of speaker recognition [12] by building a noisy model and

using it to decode noisy test segments. This iterative method compensates additive and convolutive noises directly at the data level. The main advantages of this method are to allow the compensation of the noise presenting in both test and training data, to take into account the variance of the different noises, and to facilitate the use of delta coefficients.

2.1.4 Score Normalization

A robust back-end training called "multi-style" [13] was proposed as a possible solution to noise reduction in the score level. This method used a large set of clean and noisy data (affected with different noises and SNR levels) to build a generic scoring model. The obtained model gave good performance in general but was still suboptimal (for a particular noise) because of its generalization (the same system was used for all noises). Adding noisy training data in the current i-vector based approach followed by probabilistic linear discriminant analysis (PLDA) can bring significant gains in accuracy at various signal-to-noise ratio (SNR) levels. Besides, [14] proposed a method for determining the nuisance characteristics presenting in an audio signal. The method relied on the extraction of i-vectors over the signal, an approach borrowed from the speaker recognition literature. Given a set of audio classes in the training data, a Gaussian model was trained to represent the i-vectors for each of these classes. During recognition, these models were used to obtain the posterior probability of each class given the i-vector for a certain signal. This framework allowed for a unified way of detecting any kind of nuisance characteristic that was properly encoded in the i-vector used to represent the signal.

2.2 Channel Mismatch

Channel mismatch is another salient factor that influences the recognition performance. In real applications, speech utterances are often recorded with various types of microphones (such as desktop microphone and head phone), and these speech signals are changed in some degree due to different transmission channels. In every Speaker Recognition Evaluations organized by NIST [14], the channel mismatch issue was always regarded as one of the most important challenges. To encourage the research dealing with channel mismatch issues, different recording devices and transmission channels have been utilized in collecting the evaluation data [15, 16]. Nowadays, research dealing with the channel mismatch of speaker verification tasks can be categorized into three directions: feature transformation, model compensation, and score normalization.

2.2.1 Feature Transformation

CMS (Cepstral Mean Subtraction) [17] or Cepstral Mean Normalization (CMN) which subtracts the mean value of each feature vector over the entire utterance is the simplest and most commonly-used method for many speaker verification systems. The channel variations are considered to be stable over the entire utterance in these methods. Feature mapping [18] that maps features to a channel-independent feature space and feature warping which modifies the short-term feature distribution to follow a reference distribution are also effective methods but with more complex implementation.

2.2.2 Channel Compensation

SMS (Speaker Model Synthesis) is popular in GMM-UBM systems, which transforms models from one channel to another according to the UBM deviations between channels. Reference [19] proposed a novel statistical modeling and compensation method. In channel-mismatched conditions, the new approach uses speaker-independent channel transformations to synthesize a speaker model that corresponds to the channel of the test session. A cohort-based speaker model synthesis (SMS) algorithm, designed for synthesizing robust speaker models without requiring channel-specific enrollment data, was proposed in [20]. This algorithm utilized a priori knowledge of channels extracted from speaker-specific cohort sets to synthesize such speaker models. Besides, Ref. [21] explored techniques specific to the SVM framework in order to derive fully non-linear channel compensations.

Factor analysis is another model-level compensation method to analyze the discrimination of speaker models over different channels. [22] proposed a hybrid compensation scheme (both in the feature and the model domains). The implementation is simpler, as the target speaker model does not change over the verification experiment and the standard likelihood computation can be employed. In addition, while the classical compensation scheme brings a bias in scores (score normalization is needed to obtain good performance), this approach presents good results with native scores. Finally, the use of a SVM classifier with a proper supervector-based kernel is straightforward. JFA (Joint Factor Analysis) [23], a more comprehensive statistical approach, has gained much success in speaker verification. The speaker variations and channel (session) variations were modeled as independent variables spanning in a low-rank subspace, which defined the speaker-and channel-variations as two independent random variables following a priori standard Gaussian distributions. Then the factors were inferred the posterior probability of the speaker-and channel-variations from the given speech.

The i-vector method [24] assumes that the speaker and channel variations cannot be separated by JFA because the channel variation also contains speaker information. So in the i-vector method, a low-rank total variability space is defined to

represent speaker-and channel-variations at the same time, and the speaker utterance is represented by an i-vector which is derived by inferring the posterior distribution of the total variance factor. There is no distinction between speaker effects and channel effects in GMM supervector space. Both speaker-and channel-variations are retained in i-vector. However, the total representation will lead to less discrimination among speakers due to channel variations. Therefore, many inter-channel compensation methods especially some popular discriminative approaches are employed to extract more accentuated speaker information. WCCN (With-in Class Covariance Normalization) [25] and LDA (Linear Discriminant Analysis) [24, 26] are both linear transformation to optimize the linear kernels. NAP (Nuisance Attribute Projection) [21] is to find the projection optimized by minimizing the difference among channels.

The most recent research focuses on the PLDA (Probabilistic Linear Discriminant Analysis) [27, 28], which can improve the performance of an i-vector system greatly. PLDA is a probabilistic version of LDA, and also is a generative model that utilizes a prior distribution on the speaker-and channel-variations. PLDA plus length normalization was reported to be most effective. The success of this model is believed to be largely attributed to two factors: one is its training objective function that reduces the intra-speaker variation while enlarges inter-speaker variation, and the other is the Gaussian prior that is assumed over speaker vectors, which improves robustness when inferring i-vectors for speakers with limited training data.

These two factors, however, are also two main shortcomings of the PLDA model. As for the objective function, although it encourages discrimination among speakers, the task in speaker recognition is to discriminate the true speaker and the imposters which is a binary decision rather than the multi-class discrimination in PLDA training. As for the Gaussian assumption, although it greatly simplifies the model, the assumption is rather strong and is not practical in some scenarios, leading to less representative models.

Some researchers have noticed these problems. For example, to go beyond the Gaussian assumption, Kenny [29] proposed a heavy-tailed PLDA which assumed a non-Gaussian prior over the speaker mean vector. Garcia-Romero and Espy-Wilson [30] found that length normalization could compensate for the non-Gaussian effect and boost performance of Gaussian PLDA to the level of the heavy-tailed PLDA. Burget, Cumani and colleagues [31, 32] proposed a pair-wised discriminative model discriminating the true speakers and the imposters. In their approach, the model accepted a pair of i-vectors and predicted the probability that how they belong to the same speaker. The input features of the model were derived from the i-vector pairs according to a form derived from the PLDA score function (further generalized to any symmetric score functions in [32]), and the model was trained on i-vector pairs that have been labelled as identical or different speakers. A particular shortcoming of this approach was that the feature expansion was highly complex. To solve this problem, a partial discriminative training approach was proposed in

[33], which optimized the discriminative model on a subspace without requiring any feature expansion. In [34], Wang proposed a discriminative approach based on deep neural networks (DNN), sharing the same idea as in the pair-wised training, whereas the features were defined manually.

2.2.3 Score Normalization

The score normalization algorithms include Hnorm [35, 36], Htnorm [37], Cnorm [38], and Atnorm [39], which utilize some a priori knowledge of channels to normalize impostors' verification scores into a standard normal distribution, so as to remove the influence of channel distortions from verification scores.

2.3 Multiple Speakers

Speaker recognition has generally been viewed as a problem of verifying or identifying a particular speaker in a speech segment containing only a single speaker. But for some real applications, the problem is to verify or identify particular speakers in a speech segment containing multiple speakers [40, 41]. Due to the diversity of multiple speaker speech, it is much more complex than a single speaker recognition and it requires high robustness to the existing system.

In a multiple-speaker scenario, if the system cannot separate single speaker segments effectively, it will directly affect the system performance. Automatic systems need to be able to segment the speech containing multiple speakers into segments and to determine whether the speech by a particular speaker is present and where in the segment this speech occurs [41], and then a series of single speaker recognition approaches could be performed. Figure 2.1 is a basic framework of a multiple-speaker recognition system. When multiple-speaker speech coming, the first procedure performs noise reduction to purify the speech audio. Following, the feature extraction and the speech activity detection is then normally performed to remove the influence of non-speech segments. Single speaker segments are extracted with speaker segmentation and clustering, and then recognition performs the same way as in single speaker recognition. Current research directions in multiple-speaker tasks include: robust features, robust speaker models, and segmentation and clustering algorithms. Robust features focus on extracting effective features in multiple-speaker scenarios, apart from MFCC, time-delay features,

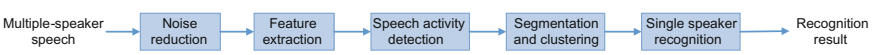

Fig. 2.1 System framework of multiple-speaker recognition

prosodic features and voice-quality features can be beneficial. Robust speaker models describe speakers in a short-time pronunciation to make more stable speaker representation. Moreover, the speaker segmentation and the clustering algorithms are two core techniques in multiple-speaker recognition. After several iterations of segmentation and clustering till convergence, the performance of multi-speaker recognition system will be improved.

2.3.1 Robust Features

One of the factors that critically affects the performance of a multiple-speaker task is the process of feature extraction. MFCC is one of the most commonly used short-term speech features in speaker recognition, and also effectively applied to multiple-speaker tasks. Besides, the time-delay feature is successful for speaker localization especially under the multiple microphone channels. Assuming that the position of any speaker does not change, the speaker localization may thus be used as alternative features in multiple-speaker tasks [42], which has become extremely popular. Some research also combined acoustic features and inter-channel delay features at the weighted log-likelihood level [43]. The use of prosodic features is emerging as a reaction to the theoretical inconsistency derived from using MFCC features for both speaker recognition and speech recognition. In [44], the authors presented a systematic investigation which showed both short-term cepstral features and long-term prosodic feature can provide significant information for speaker discrimination. References [45, 46] fused the above two features with jitter and shimmer voice-quality features and achieved considerable improvements.

2.3.2 Robust Speaker Models

One of the key points in multiple-speaker recognition is how to accurately represent speakers within a short utterance. So robust speaker modeling has become another research topic to improve the multiple-speaker recognition performance. The eigenvoice-based modeling has shown its advantages to represent speakers in speaker segmentation tasks [47], and it gets the *prior* knowledge about the speaker space to find a low dimensional vector of speaker factors that summarize the salient speaker characteristics. The speaker factors can be computed effectively in a small size window and do not suffer the problem of data sparseness. Reference [48] used an i-vector-based approach to search the speaker change points with the same idea. In order to enhance the stability and hence to improve the performance, Ref. [49] proposed a method based on a set of reference speaker models which can make a representation of the whole speaker model space. Recently, with the great success of deep neural network, Ref. [50] first proposed a novel deep neural architecture (DNN) especially for learning speaker-specific characteristics from MFCC, and

used this kind of speaker-specific characteristics to perform speaker segmentation and clustering. Reference [51] compared 4 types of neural network feature transforms and found that classification networks can achieve better result than comparison networks in multiple-speaker tasks.

2.3.3 Segmentation and Clustering Algorithms

There are two types of algorithms for speaker segmentation and clustering. One is to unity the segmentation and clustering tasks into one-step, and the other is to perform the segmentation and the clustering tasks independently. The former is to identify the speaker information while getting the speaker segments [52]; The latter is to segment the audio from multiple speakers into speech segments from single speakers, and then cluster the segments from the same speakers for independent identification.

The state-of-the-art one-step segmentation and clustering algorithm is based on E-HMM models [53]. Each of them can fit into two categories: the bottom-up [54, 55] and the top-down [56] approach. The bottom-up approach is initialized with many clusters (usually more clusters than actual speakers), and the top-down approach is initialized with one cluster. In both cases, the aim is to iteratively converge towards an optimum number of clusters. Reference [57] made a comparative study of these two approaches and concluded that the bottom-up approach can capture comparatively purer models and thus can be more sensitive to nuisance variation such as the speech content; while the top-down approach can produce less discriminative speaker models but can potentially better normalized against nuisance variation. To solve the problem against initialization parameter variation in the bottom-up approach, Ref. [58] presented a method to reduce manual tuning of these values. In the E-HMM segmentation and clustering approach, the problem is rendered particularly difficulty by the fact that there is not any *a prior* knowledge of the number of speakers. Reference [59] addressed this problem with the hierarchical Dirichlet process hidden Markov model (HDP-HMM) and sticky HDP-HMM. Reference [60] proposed a process for including priors of speaker counting with agglomerative hierarchical clustering and demonstrated significantly improvement in terms of calibration error for speaker diarization.

When performing segmentation and clustering separately, some distance measures need to be pre-defined to detect speaker change points in the segmentation step. The Bayesian Information Criterion (BIC) is one kind of method for model selection and was used as the distance measure in [61]. The Generalized Likelihood Ratio (GLR) [62] is another effective distance measure method. Reference [63] used Kullback-Leibler Divergence (KL) to perform speaker segmentation in broadcast news tasks and achieved great success. Support Vector Machine (SVM) is a type of expeditious classification and was used for segmentation in [64]. In the clustering step, in order to reduce the cost of computation, the agglomerative information bottleneck (aIB) [65] and the sequential information bottleneck (sIB) [66] were

proposed and they therefore were widely used in the meeting scenarios. Reference [67] made a noise-robust speaker clustering based on spectral clustering and compared it with the hierarchical clustering and the K-means clustering. Reference [68] proposed a novel DNN-based clustering and performed re-segmentation and clustering in each iteration. Reference [69] investigated how accurate the clustering algorithm will be depending on the characteristics of the audio stream, which was an effective guidance of speaker clustering. ELISA [70] was a hybrid system combining the segmentation and the clustering steps together.

2.4 Discussions

Background noise, channel mismatch and multiple speakers are three most common factors that will influence the performance of speaker recognition systems. In real applications, there is often *no prior* knowledge of environmental noise, transmission channel and number of speakers containing in the speech segment in advance. Therefore, it is difficult to pre-train the noise/channel model and define the clustering number. To deal with these environment-related issues, researchers have carried to do some studies from different of views. In this chapter, we summarize the latest research studies and techniques among these three factors from different aspects. We believe that these three factors are still the main research directions.

References

1. Boll S (1979) Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans Acoust Speech Signal Process 27(2):113–120
2. Wang N, Ching PC, Zheng N et al (2011) Robust speaker recognition using denoised vocal source and vocal tract features. IEEE Trans Audio Speech Lang Process 19(1):196–205
3. Hermansky H, Morgan N (1994) RASTA processing of speech. IEEE Trans Speech Audio Process 2(4):578–589
4. Furui S (1981) Cepstral analysis technique for automatic speaker verification. IEEE Trans Acoust Speech Signal Process 29(2):254–272
5. Zhao X, Shao Y, Wang DL (2012) CASA-based robust speaker identification. IEEE Trans Audio Speech Lang Process 20(5):1608–1616
6. Sadjadi SO, Hansen JHL (2010) Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions. INTERSPEECH, pp 2138–2141
7. Hanilçi C, Kinnunen T, Saeidi R et al (2012) Comparing spectrum estimators in speaker verification under additive noise degradation. Acoustics, speech and signal processing (ICASSP), 2012 IEEE international conference on. IEEE, pp 4769–4772
8. Lei Y, Burget L, Scheffer N (2013) A noise robust i-vector extractor using vector taylor series for speaker recognition. Acoustics, speech and signal processing (ICASSP), 2013 IEEE international conference on. IEEE, pp 6788–6791

9. Lei Y, McLaren M, Ferrer L et al (2014) Simplified vts-based i-vector extraction in noise-robust speaker recognition. Acoustics, speech and signal processing (ICASSP), 2014 IEEE international conference on. IEEE, pp 4037–4041
10. Martinez D, Burget L, Stafylakis T et al (2014) Unscented transform for ivector-based noisy speaker recognition. Acoustics, speech and signal processing (ICASSP), 2014 IEEE international conference on. IEEE, pp 4042–4046
11. Gales MJF, Young SJ (1996) Robust continuous speech recognition using parallel model combination. IEEE Trans Speech Audio Process 4(5):352–359
12. Bellot O, Matrouf D, Merlin T et al (2000) Additive and convolutional noises compensation for speaker recognition. INTERSPEECH, pp 799–802
13. Lei Y, Burget L, Ferrer L et al (2012) Towards noise-robust speaker recognition using probabilistic linear discriminant analysis. Acoustics, speech and signal processing (ICASSP), 2012 IEEE international conference on. IEEE, pp 4253–4256
14. Doddington GR, Przybocki MA, Martin AF et al (2000) The NIST speaker recognition evaluation–overview, methodology, systems, results, perspective. Speech Commun 31 (2):225–254
15. The NIST year 2012 speaker recognition evaluation plan. https://www.nist.gov/sites/default/files/documents/itl/iad/mig/NIST_SRE12_evalplan-v17-r1.pdf
16. NIST 2016 speaker recognition evaluation plan. https://www.nist.gov/sites/default/files/documents/itl/iad/mig/SRE16_Eval_Plan_V1-0.pdf
17. Furui S (1981) Cepstral analysis technique for automatic speaker verification. IEEE Trans Acoust Speech Signal Process 29(2):254–272
18. Reynolds DA (2003) Channel robust speaker verification via feature mapping. Acoustics, speech, and signal processing, 2003. Proceedings. (ICASSP'03). 2003 IEEE international conference on. IEEE, vol 2, pp 2–53
19. Teunen R, Shahshahani B, Heck LP (2000) A model-based transformational approach to robust speaker recognition. INTERSPEECH, pp 495–498
20. Wu W, Zheng TF, Xu MX et al (2007) A cohort-based speaker model synthesis for mismatched channels in speaker verification. IEEE Trans Audio Speech Lang Process 15 (6):1893–1903
21. Solomonoff A, Quillen C, Campbell WM (2004) Channel compensation for SVM speaker recognition. Odyssey, vol 4, pp 219–226
22. Matrouf D, Scheffer N, Fauve BGB et al (2007) A straightforward and efficient implementation of the factor analysis model for speaker verification. INTERSPEECH, pp 1242–1245
23. Kenny P, Boulianne G, Ouellet P et al (2007) Joint factor analysis versus eigenchannels in speaker recognition. IEEE Trans Audio Speech Lang Process 15(4):1435–1447
24. Dehak N, Kenny PJ, Dehak R et al (2011) Front-end factor analysis for speaker verification. IEEE Trans Audio Speech Lang Process 19(4):788–798
25. Hatch AO, Kajarekar SS, Stolcke A (2006) Within-class covariance normalization for SVM-based speaker recognition. INTERSPEECH
26. McLaren M, Van Leeuwen D (2011) Source-normalised-and-weighted LDA for robust speaker recognition using i-vectors. Acoustics, speech and signal processing (ICASSP), 2011 IEEE international conference on. IEEE, pp 5456–5459
27. Ioffe S (2006) Probabilistic linear discriminant analysis. European conference on computer vision. Springer, Berlin, pp 531–542
28. Prince SJD, Elder JH (2007) Probabilistic linear discriminant analysis for inferences about identity. Computer vision, 2007. ICCV 2007. IEEE 11th international conference on. IEEE, pp 1–8
29. Kenny P (2010) Bayesian speaker verification with heavy-tailed priors. Odyssey, pp 14
30. Garcia-Romero D, Espy-Wilson CY (2011) Analysis of i-vector length normalization in speaker recognition systems. INTERSPEECH, pp 249–252

31. Burget L, Plchot O, Cumani S et al (2011) Discriminatively trained probabilistic linear discriminant analysis for speaker verification. Acoustics, speech and signal processing (ICASSP), 2011 IEEE international conference on. IEEE, pp 4832–4835
32. Cumani S, Brummer N, Burget L et al (2013) Pairwise discriminative speaker verification in the i-vector space. IEEE Trans Audio Speech Lang Process 21(6):1217–1227
33. Hirano I, Lee KA, Zhang Z et al (2014) Single-sided approach to discriminative PLDA training for text-independent speaker verification without using expanded i-vector. Chinese spoken language processing (ISCSLP), 2014 9th international symposium on. IEEE, pp 59–63
34. Wang J, Wang D, Zhu Z et al (2014) Discriminative scoring for speaker recognition based on i-vectors. Asia-pacific signal and information processing association, 2014 annual summit and conference (APSIPA). IEEE, pp 1–5
35. Reynolds DA, Quatieri TF, Dunn RB (2000) Speaker verification using adapted Gaussian mixture models. Digit Signal Proc 10(1–3):19–41
36. Reynolds DA (1997) Comparison of background normalization methods for text-independent speaker verification. Eurospeech
37. Auckenthaler R, Carey M, Lloyd-Thomas H (2000) Score normalization for text-independent speaker verification systems. Digit Signal Proc 10(1):42–54
38. Bimbot F, Bonastre JF, Fredouille C et al (2004) A tutorial on text-independent speaker verification. EURASIP J Appl Sig Process 2004:430–451
39. Sturim DE, Reynolds DA (2005) Speaker adaptive cohort selection for Tnorm in text-independent speaker verification. Acoustics, speech, and signal processing, 2005. Proceedings. (ICASSP'05). IEEE international conference on. IEEE, 1: I/741-I/744 vol 1
40. Anguera X, Bozonnet S, Evans N et al (2012) Speaker diarization: a review of recent research. IEEE Trans Audio Speech Lang Process 20(2):356–370
41. Martin AF, Przybocki MA (2001) Speaker recognition in a multi-speaker environment. INTERSPEECH, pp 787–790
42. Lathoud G, McCowan IA (2003) Location based speaker segmentation. Multimedia and expo, 2003. ICME'03. Proceedings. 2003 international conference on. IEEE, vol 3, pp 3–621
43. Pardo JM, Anguera X, Wooters C Speaker diarization for multiple distant microphone meetings: mixing acoustic features and inter-channel time differences. INTERSPEECH
44. Friedland G, Vinyals O, Huang Y et al (2009) Prosodic and other long-term features for speaker diarization. IEEE Trans Audio Speech Lang Process 17(5):985–993
45. Woubie A, Luque J, Hernando J (2015) Using voice-quality measurements with prosodic and spectral features for speaker diarization. INTERSPEECH, pp 3100–3104
46. Woubie A, Luque J, Hernando J (2016) Short-and long-term speech features for hybrid HMM-i-Vector based speaker diarization system. Odyssey
47. Castaldo F, Colibro D, Dalmasso E et al (2008) Stream-based speaker segmentation using speaker factors and eigenvoice. Acoustics, speech and signal processing, 2008. ICASSP 2008. IEEE international conference on. IEEE, pp 4133–4136
48. Desplanques B, Demuynck K, Martens JP (2015) Factor analysis for speaker segmentation and improved speaker diarization. INTERSPEECH. Abstracts and proceedings USB productions, pp 3081–3085
49. Wang G, Zheng TF (2009) Speaker segmentation based on between-window correlation over speakers' characteristics. Proceedings: APSIPA ASC, pp 817–820
50. Chen K, Salman A (2011) Learning speaker-specific characteristics with a deep neural architecture. IEEE Trans Neural Networks 22(11):1744–1756
51. Yella SH, Stolcke A (2015) A comparison of neural network feature transforms for speaker diarization. INTERSPEECH, pp 3026–3030
52. Kotti M, Moschou V, Kotropoulos C (2008) Speaker segmentation and clustering. Sig Process 88(5):1091–1124
53. Meignier S, Bonastre JF, Igounet S (2001) E-HMM approach for learning and adapting sound models for speaker indexing. A speaker Odyssey-the speaker recognition workshop

54. Meignier S, Bonastre JF, Fredouille C et al (2000) Evolutive HMM for multi-speaker tracking system. Acoustics, speech, and signal processing, 2000. ICASSP'00. Proceedings. 2000 IEEE international conference on. IEEE, vol 2, pp 1201–1204
55. Ajmera J, Wooters C (2003) A robust speaker clustering algorithm. Automatic speech recognition and understanding, 2003. ASRU'03. 2003 IEEE Workshop on. IEEE, pp 411–416
56. Wooters C, Huijbregts M (2008) The ICSI RT07s speaker diarization system. Multimodal technologies for perception of humans. Springer, Berlin, pp 509–519
57. Evans N, Bozonnet S, Wang D et al (2012) A comparative study of bottom-up and top-down approaches to speaker diarization. IEEE Trans Audio Speech Lang Process 20(2):382–392
58. Imseng D, Friedland G (2010) Tuning-robust initialization methods for speaker diarization. IEEE Trans Audio Speech Lang Process 18(8):2028–2037
59. Fox EB, Sudderth EB, Jordan MI et al (2011) A sticky HDP-HMM with application to speaker diarization. The annals of applied statistics, pp 1020–1056
60. Sell G, McCree A, Garcia-Romero D (2016) Priors for speaker counting and diarization with AHC. INTERSPEECH 2016, pp 2194–2198
61. Chen S, Gopalakrishnan P (1998) Speaker, environment and channel change detection and clustering via the bayesian information criterion. Proc. DARPA broadcast news transcription and understanding workshop, vol 8, pp 127–132
62. Gish H, Siu MH, Rohlicek R (1991) Segregation of speakers for speech recognition and speaker identification. Acoustics, speech, and signal processing, 1991. ICASSP-91, 1991 international conference on. IEEE, pp 873–876
63. Siegler MA, Jain U, Raj B et al (1997) Automatic segmentation, classification and clustering of broadcast news audio. Proc. DARPA speech recognition workshop. 1997
64. Fergani B, Davy M, Houacine A (2008) Speaker diarization using one-class support vector machines. Speech Commun 50(5):355–365
65. Vijayasenan D, Valente F, Bourlard H (2007) Agglomerative information bottleneck for speaker diarization of meetings data. Automatic speech recognition and understanding, 2007. ASRU. IEEE workshop on. IEEE, pp 250–255
66. Vijayasenan D, Valente F, Bourlard H (2008) Combination of agglomerative and sequential clustering for speaker diarization. Acoustics, speech and signal processing, 2008. ICASSP 2008. IEEE international conference on. IEEE, pp 4361–4364
67. Tawara N, Ogawa T, Kobayashi T (2015) A comparative study of spectral clustering for i-vector-based speaker clustering under noisy conditions. Acoustics, speech and signal processing (ICASSP), 2015 IEEE international conference on. IEEE, pp 2041–2045
68. Milner R, Hain T (2016) DNN-based speaker clustering for speaker diarisation. Proceedings of the annual conference of the international speech communication association, INTERSPEECH. Sheffield, pp 2185–2189
69. Prieto JJ, Vaquero C, García P (2016) Analysis of the impact of the audio database characteristics in the accuracy of a speaker clustering system. Odyssey, pp 393–399
70. Moraru D, Meignier S, Fredouille C et al (2004) The ELISA consortium approaches in broadcast news speaker segmentation during the NIST 2003 rich transcription evaluation. Acoustics, speech, and signal processing, 2004. Proceedings. (ICASSP'04). IEEE international conference on. IEEE, vol 1, pp 1–373

http://www.springer.com/978-981-10-3237-0

Robustness-Related Issues in Speaker Recognition

Zheng, Th.F.; Li, L.

2017, X, 49 p. 12 illus., 7 illus. in color., Softcover
ISBN: 978-981-10-3237-0